(Resource-Constraint + Privacy)-Aware Data Structures Tackling Problems in Bioinformatics

クップル ドミニク
(Dominik KOEPPL)
東京医科歯科大学

公募研究
B01 班

タイトルの説明

メモリー制限 (resource constraint)
- スパコンといった効率よいハードの代わりに、携帯品を使う

機密情報 (privacy)
- 個人的なデータを隠す

用途
- バイオインフォマティクス (B04 班 : 渋谷先生の専門)

研究のキーワード :
1) 圧縮
2) 索引
3) 情報保護

私の背景

- ‘15~‘18: 博士課程　@ TU Dortmund【ドイツ】
- ‘18~’20: PostDoc @　九州大学
  上司：稲永俊介　（B01 班：竹田研究室）
- ‘20~:　助教　@　東京医科歯科大学
  上司：坂内英夫

研究の履歴

- 可逆圧縮：tudocomp
  - 圧縮方法を高速に作ったり
    圧縮の比較ができる　framework
- 文字列組合せ論
- 少領域データ構造

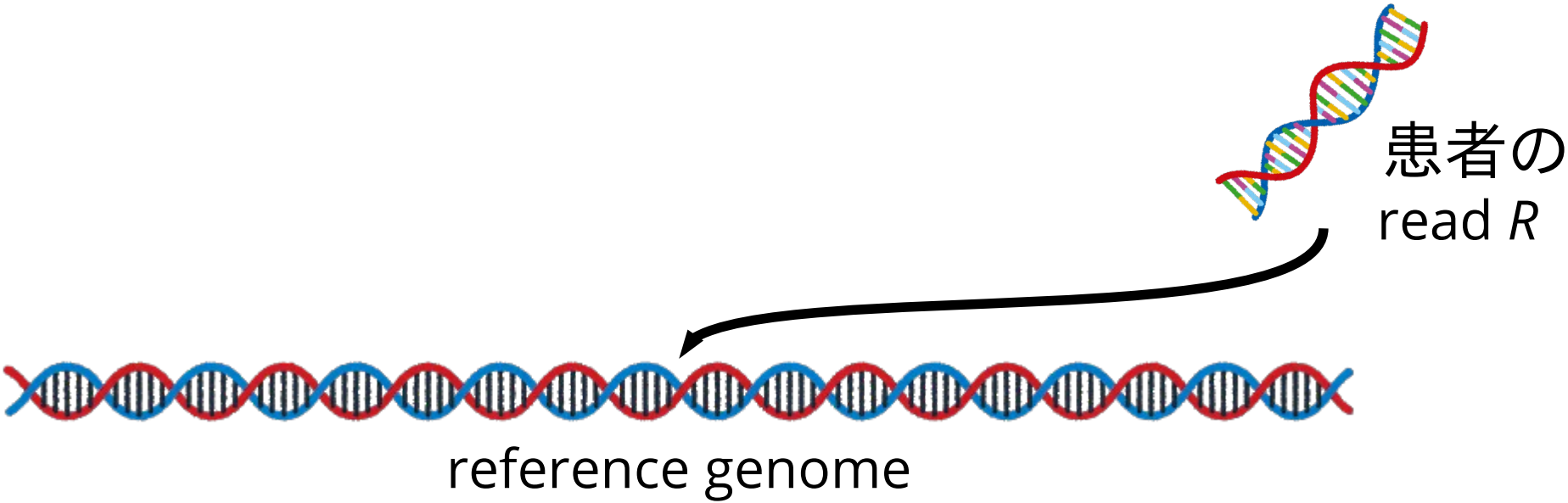
研究の背景： precision medicine

- 患者の遺伝子に基づく、特化した医療の選択
- ヒトの遺伝子 ：$3 \cdot 10^9$ 塩基対
- 患者の遺伝子の特徴をどう見つけるか？

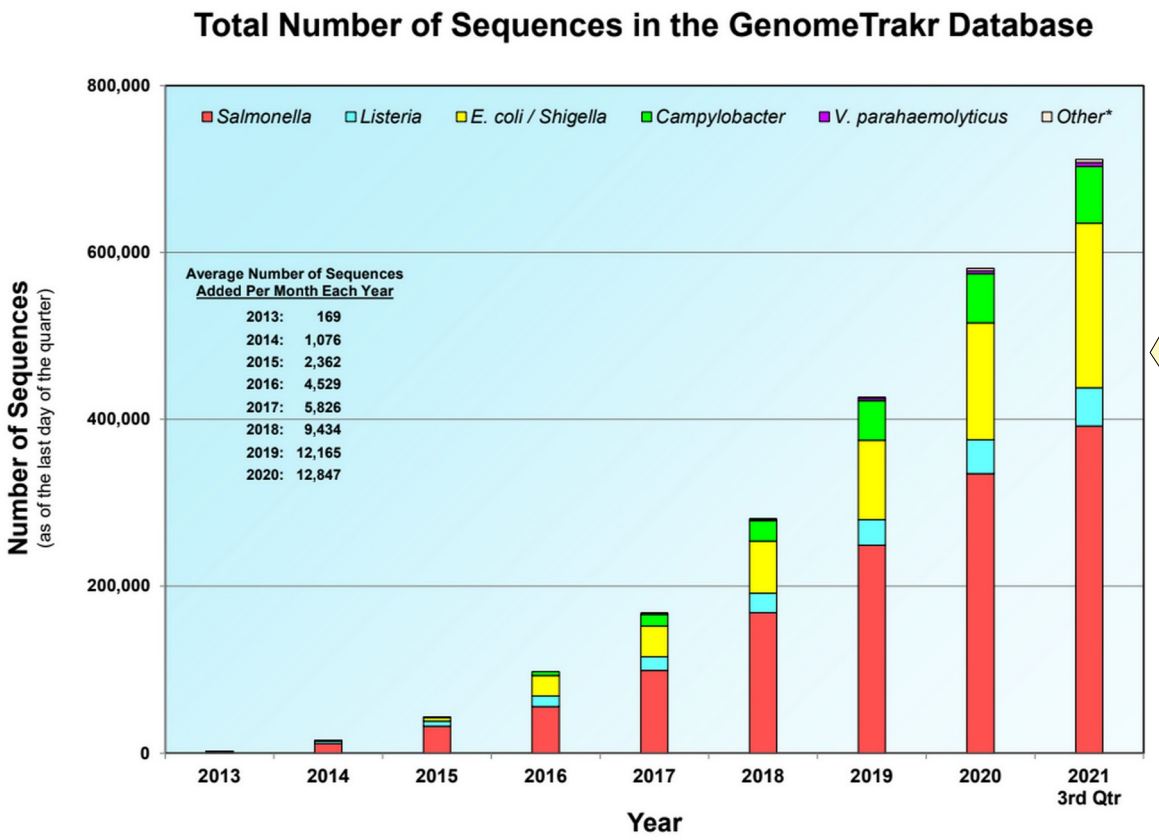
→ 患者の遺伝子が reference genome の中に出現するか検索する

read mapping

任意の read (遺伝子の部分) *R* の出現を
reference genome の中から発見する



遺伝子データの急激な増加



First sequences uploaded in February 2013

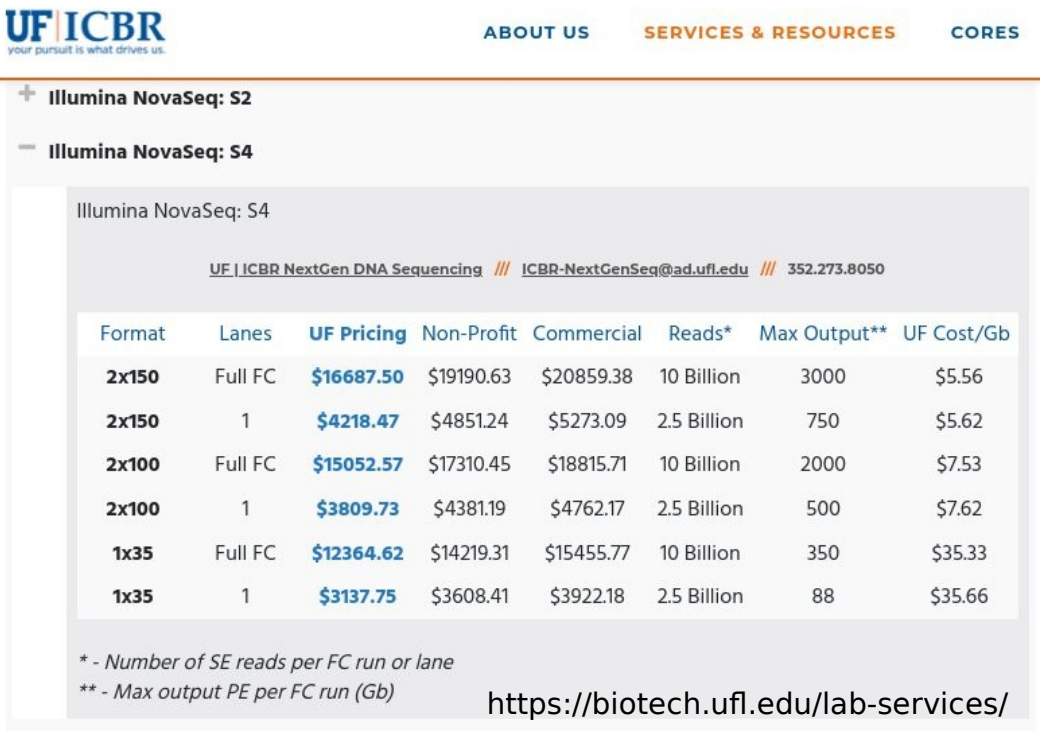
* Other pathogens: *Cronobacter, V. vulnificus, C. botulinum, and C. perfringens*

https://fda.gov

これまでに
羅列した菌の
遺伝子データが
加速的に増加している

近年の遺伝子データの羅列

- 依頼できる量が大きくなる
- 安くなる
- 保存のために
  圧縮が必須になる



UF | ICBR
your pursuit is what drives us.

ABOUT US　SERVICES & RESOURCES　CORES

+ Illumina NovaSeq: S2

− Illumina NovaSeq: S4

Illumina NovaSeq: S4

UF | ICBR NextGen DNA Sequencing /// ICBR-NextGenSeq@ad.ufl.edu /// 352.273.8050

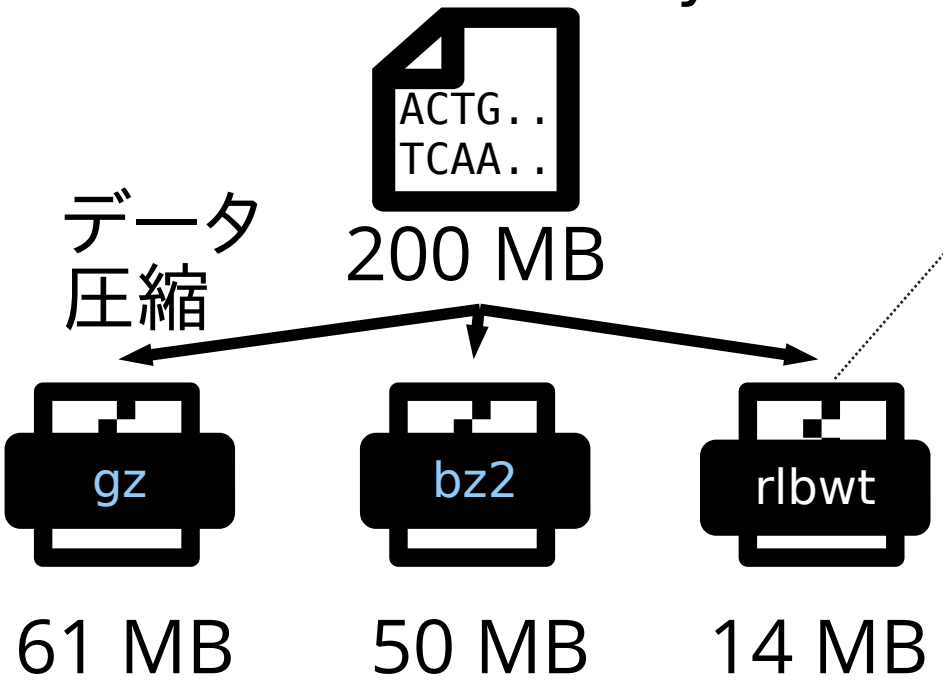
Format	Lanes	UF Pricing	Non-Profit	Commercial	Reads*	Max Output**	UF Cost/Gb
2x150	Full FC	**$16687.50**	$19190.63	$20859.38	10 Billion	3000	$5.56
2x150	1	**$4218.47**	$4851.24	$5273.09	2.5 Billion	750	$5.62
2x100	Full FC	**$15052.57**	$17310.45	$18815.71	10 Billion	2000	$7.53
2x100	1	**$3809.73**	$4381.19	$4762.17	2.5 Billion	500	$7.62
1x35	Full FC	**$12364.62**	$14219.31	$15455.77	10 Billion	350	$35.33
1x35	1	**$3137.75**	$3608.41	$3922.18	2.5 Billion	88	$35.66

** - Number of SE reads per FC run or lane*
*** - Max output PE per FC run (Gb)*

https://biotech.ufl.edu/lab-services/

1 番目の目的：圧縮

例：
入力は　Saccharomyces Cerevisiae 20 個体の DNA 列



素朴な圧縮方法なのに
圧縮率は高い

高反復データに対して、
gz・bz2 のような既存の
圧縮方法は良くない

[Dinklage ら '17]

２番目の目的：索引

Ballouz *et al. Genome Biology* (2019) 20:159
https://doi.org/10.1186/s13059-019-1774-4

Genome Biology

OPINION Open Access

Is it time to change the reference genome?

Sara Ballouz, Alexander Dobin and Jesse A. Gillis*

Abstract

The use of the human reference genome has shaped methods and data across modern genomics. This has offered many benefits while creating a few constraints. In the following opinion, we outline the history, properties, and pitfalls of the current human reference genome. In a few illustrative analyses, we focus on its use for variant-calling, highlighting its nearness to a 'type specimen'. We suggest that switching to a consensus reference would offer important advantages over the continued use of the current reference with few disadvantages.

複数の
reference genome を
利用したほうが良い

複数の reference genomes

pangenome は複数の遺伝子を纏める

⇒ 入力のデータは膨大

- 圧縮が求められるが、検索は困難になる

Briefings in Bioinformatics, 19(1), 2018, 118–135

doi: 10.1093/bib/bbw089
Advance Access Publication Date: 21 October 2016
Paper

Computational pan-genomics: status, promises and challenges

The Computational Pan-Genomics Consortium*

Abstract

Many disciplines, from human genetics and oncology to plant breeding, microbiology and virology, commonly face the challenge of analyzing rapidly increasing numbers of genomes. In case of *Homo sapiens*, the number of sequenced genomes will approach hundreds of thousands in the next few years. Simply scaling up established bioinformatics pipelines will not be sufficient for leveraging the full potential of such rich genomic data sets. Instead, novel, qualitatively different computational methods and paradigms are needed. We will witness the rapid extension of computational pan-genomics, a new sub-area of research in computational biology. In this article, we generalize existing definitions and understand a pan-genome as any collection of genomic sequences to be analyzed jointly or to be used as a reference. We examine already available approaches to construct and use pan-genomes, discuss the potential benefits of future technologies and methodologies and review open challenges from the vantage point of the above-mentioned biological disciplines. As a prominent example for a computational paradigm shift, we particularly highlight the transition from the representation of reference genomes as strings to representations as graphs. We outline how this and other challenges from different application domains translate into common computational problems, point out relevant bioinformatics techniques and identify open problems in computer science. With this review, we aim to increase awareness that a joint approach to computational pan-genomics can help address many of the problems currently faced in various domains.

Key words: pan-genome; sequence graph; read mapping; haplotypes; data structures

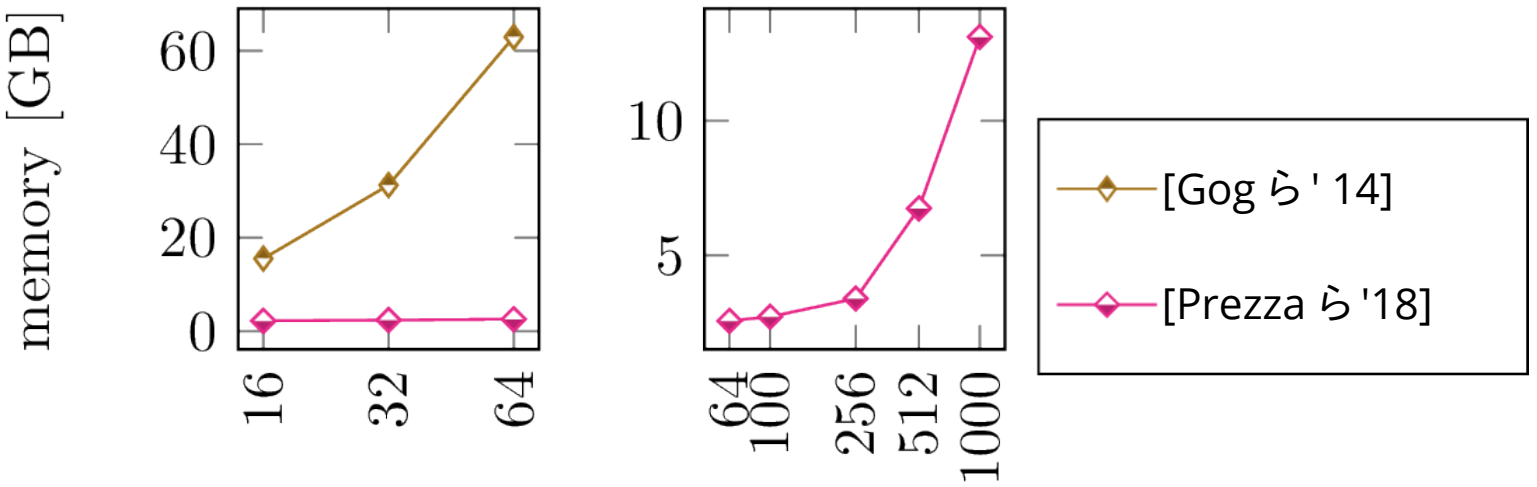
圧縮だけではなく、既存のデータ扱い方は不十分だ

pangenome に対するデータ構造が必要だ

解決策： compressed index

- 圧縮したデータの上に索引を作ることができる（例：文法圧縮、 rlbwt 、など）
- しかし構築はたくさん領域が必要

索引構築領域の増加



ヒトの染色体 19 番目の個体数

比較:
[Boucher ら '21]

２番目の目的：索引

[Gog ら '14]

圧縮された接尾辞木

- たくさん領域が必要
- 機能：豊富

[Prezza ら '18]

rlbwt に基づいて索引

- 少領域
- 機能：少ない

同じような機能を実現するために
新しいアルゴリズムの提案が必要

3 番めの目的：情報保護

医療データは個人的な情報なので、漏洩は厳禁

- Datafly system [Sweeney '98]: 住所などを隠す
- しかし、遺伝子だけでも、個人が特定できる
  (虹彩の色といった対立遺伝子から容姿を
  特定できる）[Bonomi ら '20]

⇒ pangenome 索引から個人情報が漏洩する可能性

暗号化された BWT

- scrambled BWT [Külekci '12]
- alphabet の順列 π を無作為に作る
  - BWT($\pi(T)$) を保存し、
  - π を暗号の key にする

組合せ爆発の作り方

- key の組合せの数は σ! 種類
  DNA: σ =4, IUPAC = 16
- σ は小さい場合は、 すべての σ! 通りを簡単に計算できる (しらみつぶし探索)
- idea : 圧縮方法を合成する

$$\mathrm{BWT}(\pi(T)) \rightarrow \mathrm{BWT}(\pi(\text{文法圧縮}(T)))$$

文法の非終端記号の種類は文字より圧倒的に多い

BWT と文法の組合せ

- しかし、圧縮された文字列に基づいた BWT 索引は考えられていなかった
  - WCTA’21 (先週): [鄧 , 楷 , K., 定兼] 新しい文法圧縮が考案された
- まだ理論的に面白い結果が出ていない

B04 班

暗号以外の保護方法

- k anonymity [Sweeney, Samarati '98]
- ℓ diversity [Machanavajjhala ら '07]
- string sanitization [三重野ら '21] など

B01 班

まとめ

膨大な遺伝子データの管理

- 圧縮されたデータの索引付け
- 同時に、情報を保護する
- どこでも利用するために、少領域を目指す

これからよろしくお願い致します